

# ALCF & Blue Gene Consortium WORKSHOP SERIES



## Blue Gene/P Architecture

*Rajiv Bendale* : [bendale@us.ibm.com](mailto:bendale@us.ibm.com)

*Kirk Jordan* : [kjordan@us.ibm.com](mailto:kjordan@us.ibm.com)

*Jerrold Heyman* : [jheyman@us.ibm.com](mailto:jheyman@us.ibm.com)

*Carlos P Sosa* : [cpsosa@us.ibm.com](mailto:cpsosa@us.ibm.com)

*Bob Walkup* : [walkup@us.ibm.com](mailto:walkup@us.ibm.com)

IBM, USA



# IBM's Blue Gene<sup>®</sup> Supercomputer

## Applying *Innovation that Matters* to High Performance Computing

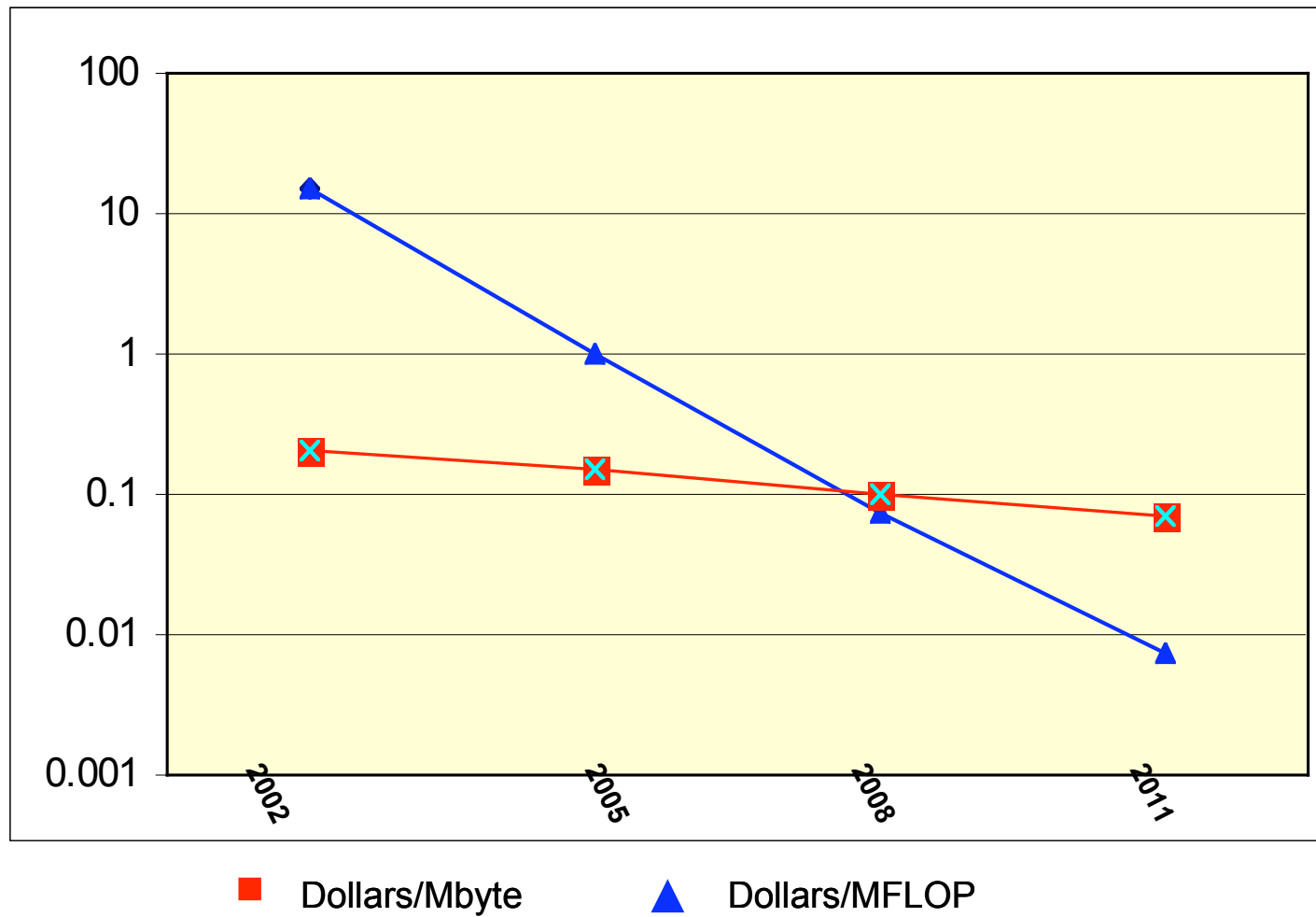
- Leadership performance, ultra scalability
- Space saving design
- Power efficient computing
- High reliability
- Ease of system management
- Familiar programming methods



# Blue Gene/P Architectural Highlights

- Scaled performance relative to BG/L through density and frequency
  - 1.2x from frequency bump 700 MHz => 850 MHz
  - 2x performance through doubling the processors/node
- Enhanced function
  - 4 way SMP, cache coherent, supports threads, OpenMP
  - Improved memory subsystem supporting higher bandwidths
  - DMA for torus, remote put-get, user programmable memory prefetch
  - Memory chip kill implemented.
  - Enhanced performance counters (including 450 core)
  - Architectures of double Hummer FPU, torus, collective network, barrier, and JTAG networks were left intact.
- Higher signaling rate
  - 2.4x higher bandwidth, lower latency for Torus and Tree networks
  - 10x higher bandwidth for Ethernet IO

## GFLOPs vs DRAM Price Reductions



# Blue Gene/P

System



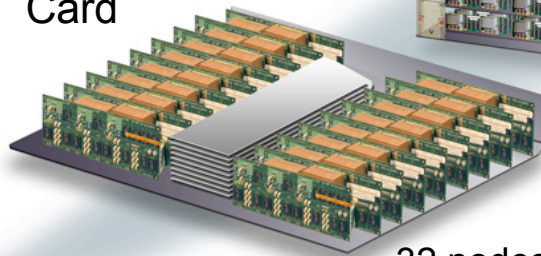
72 Racks  
73728 nodes  
294912 cores  
1 PF  
144 TB DDR2

Rack



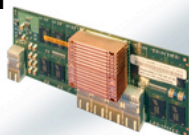
32 node-cards  
1024 nodes  
4096 cores  
13.9 TF  
2 TB DDR2

Node Card



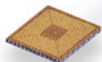
32 nodes  
128 cores  
435 GF  
64GB DDR2

Compute Card



1 node  
4 cores  
13.6 GF  
2GB DDR2

Chip



4 cores  
13.6 GF

## Blue Gene/P Packaging Characteristics

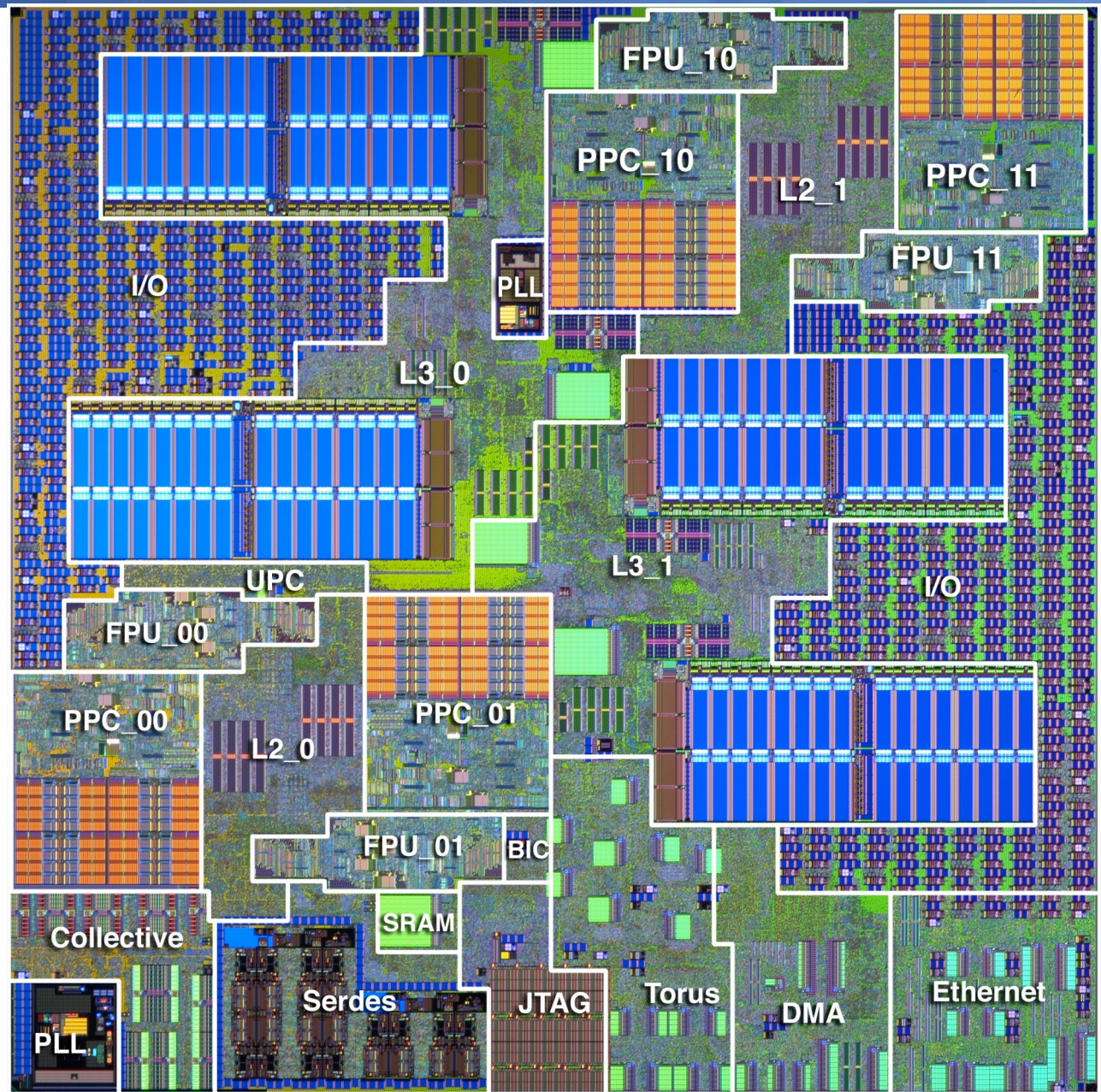
Frequency	<b>850MHz PowerPC 450</b>
First level Package	<b>FC-PBGA, 29 mm × 29 mm; 528s + 255p/g</b>
Compute card	<b>6s-8p FR4 card, 145 mm x 54.6 mm</b>
Node card	<b>6s-10p FR4 card, 427 mm x 400 mm</b>
Midplane	<b>5s-9p FR4 card, 640 mm x 550 mm</b>
Rack	<b>4ft x 3ft x 42U, 5500CFM, 40KW delivered</b>

<b>Application</b>	<b>kW within rack</b>	<b>bulk power efficiency</b>	<b>Total kW Rack Power (480VAC Wall Power)</b>
LINPACK	<b>28.7</b>	<b>0.91</b>	<b>31.5</b>
Ave application	<b>21.3</b>	<b>0.91</b>	<b>23.4</b>
Rack Idle	<b>8.60</b>	<b>0.87</b>	<b>9.9</b>



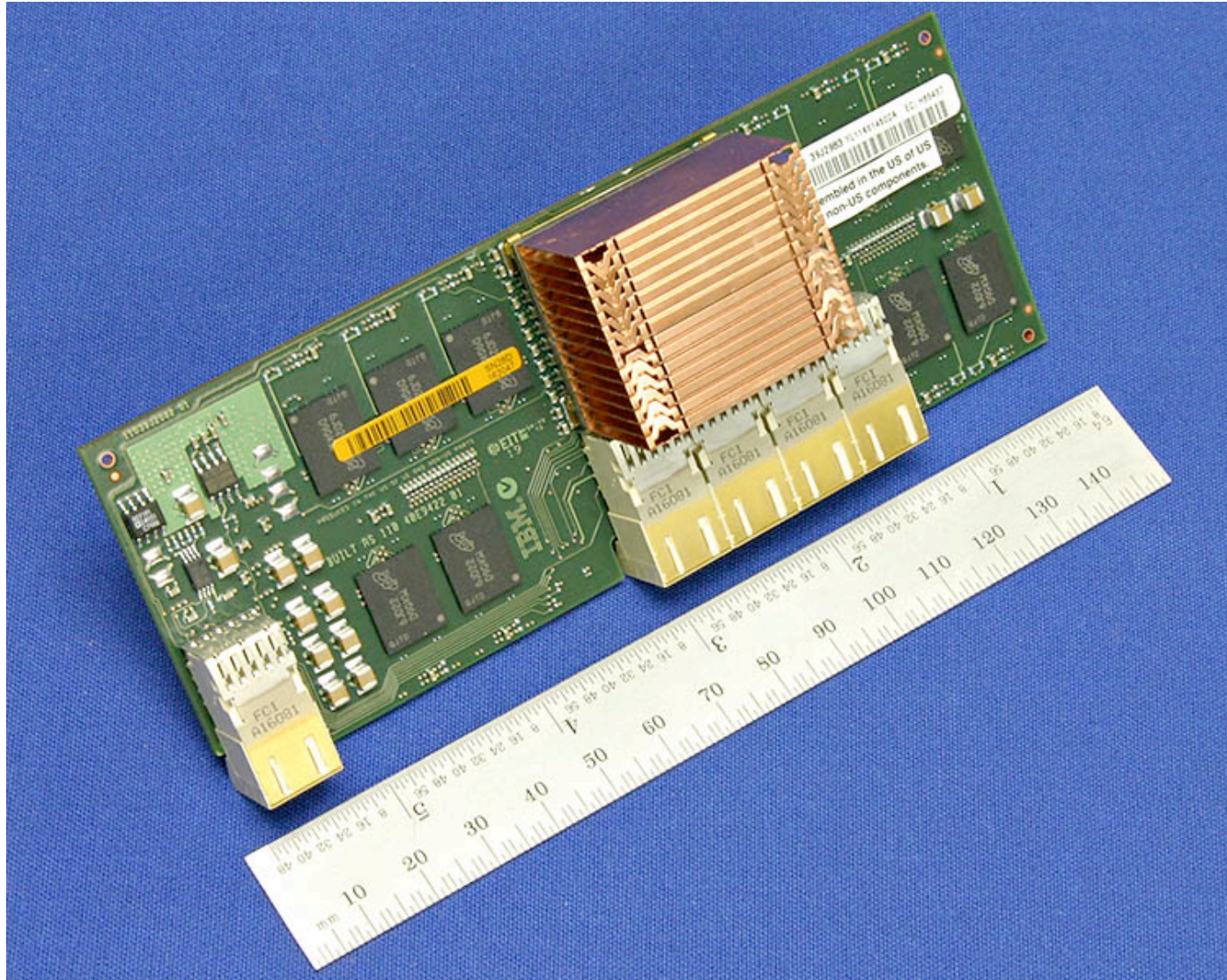
## BPC chip DD2.1 die photograph

13mmx13mm  
90 nm process  
208M transistors  
88M in eDRAM



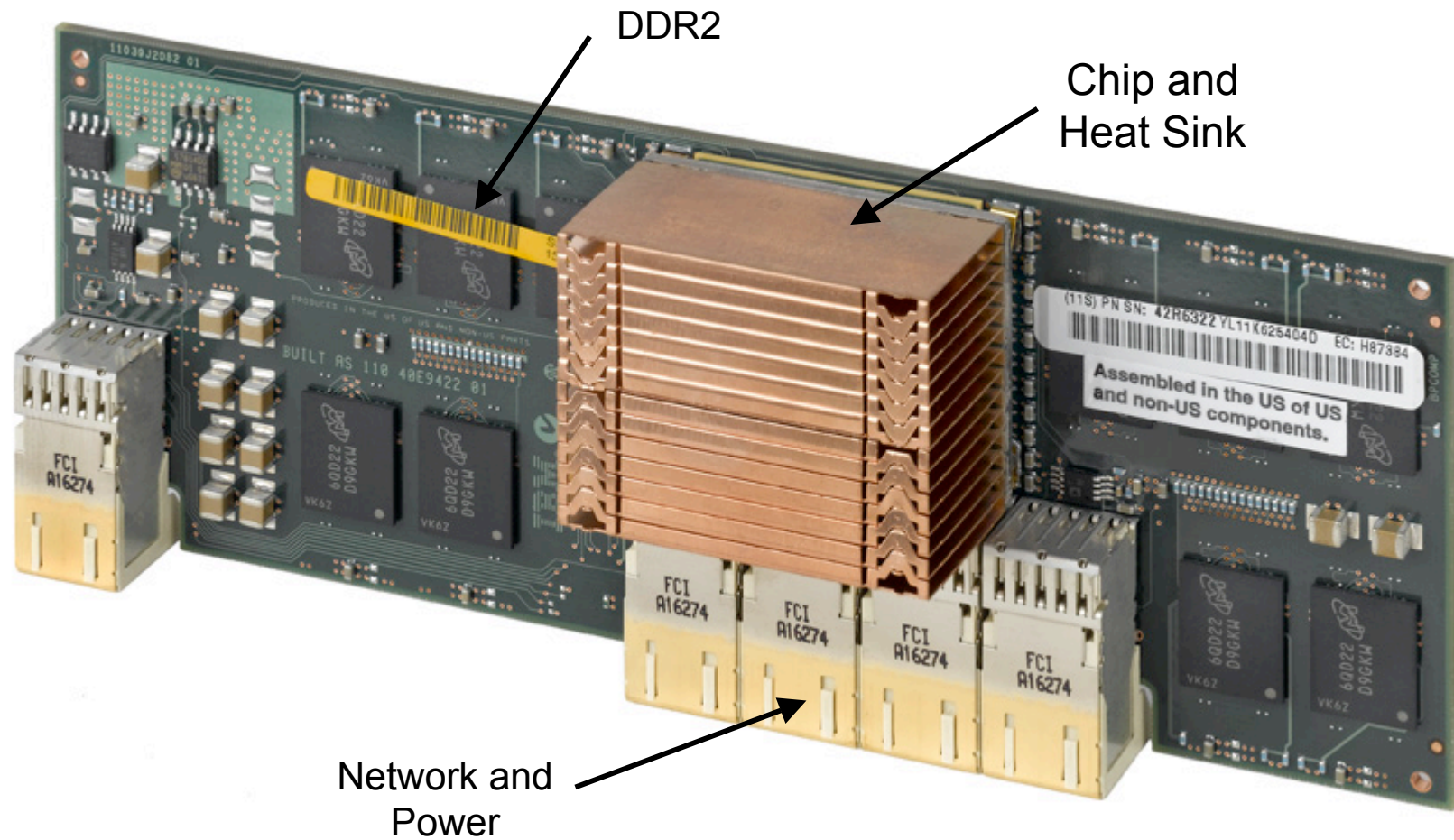


## BG/P Compute Card



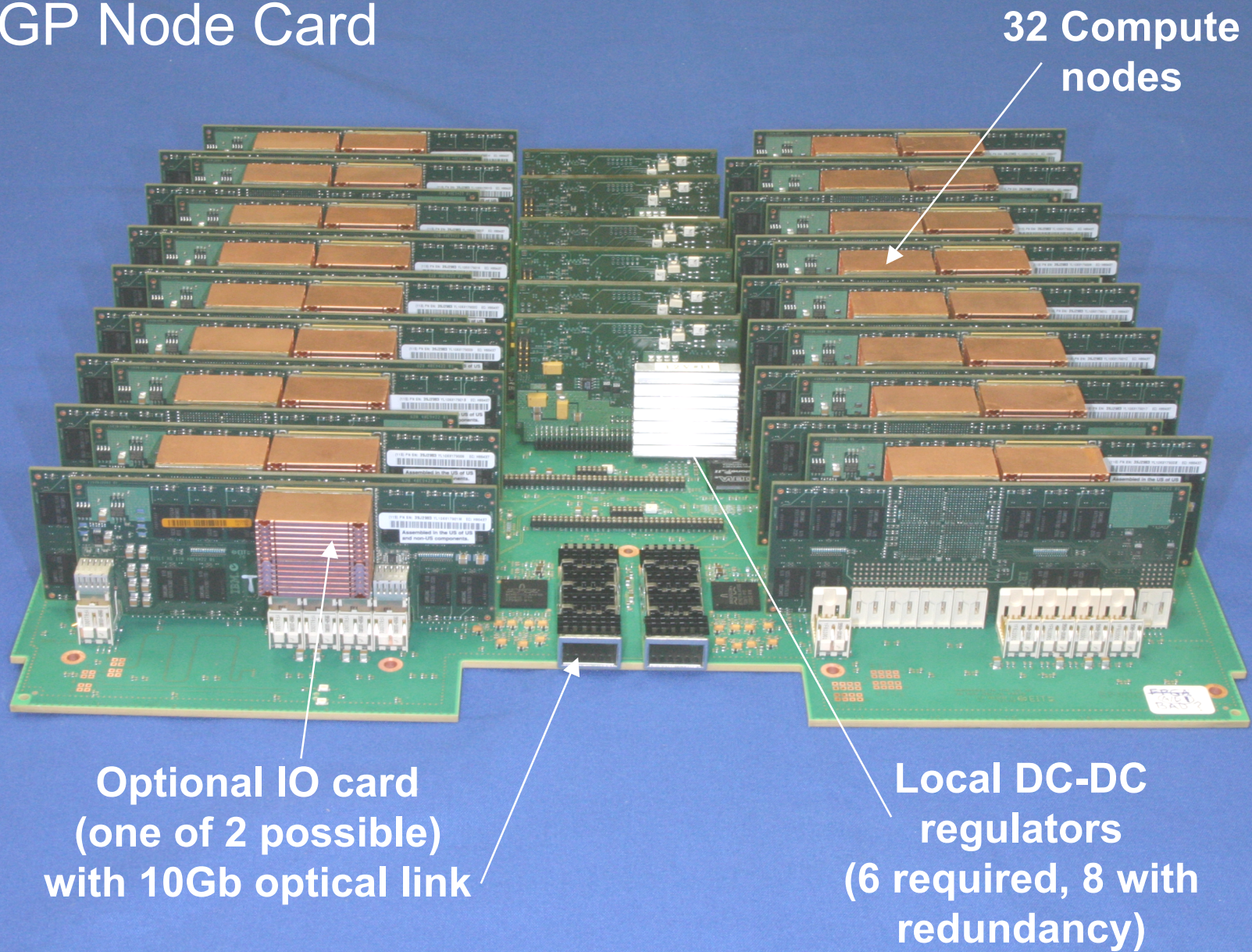


## BG/P Compute Card





# BGP Node Card



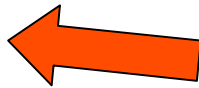


# 32 Compute Nodes

128 cores

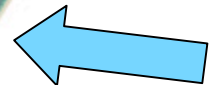
Hottest ASIC T<sub>j</sub>  
80°C@24W, 55°C@15W

Outlet Air



Max +10°C

Inlet Air



min 2.5m/s  
max 17°C

Hottest DRAM  
T<sub>case</sub> 75°C@0.3W

Optional IO card  
(1 of 2 possible)

10Gb Ethernet

Local 48V input DC-DC regulators  
5+1, 3+1 with redundancy.  
Vicor technology, tcase  
60°C@120A

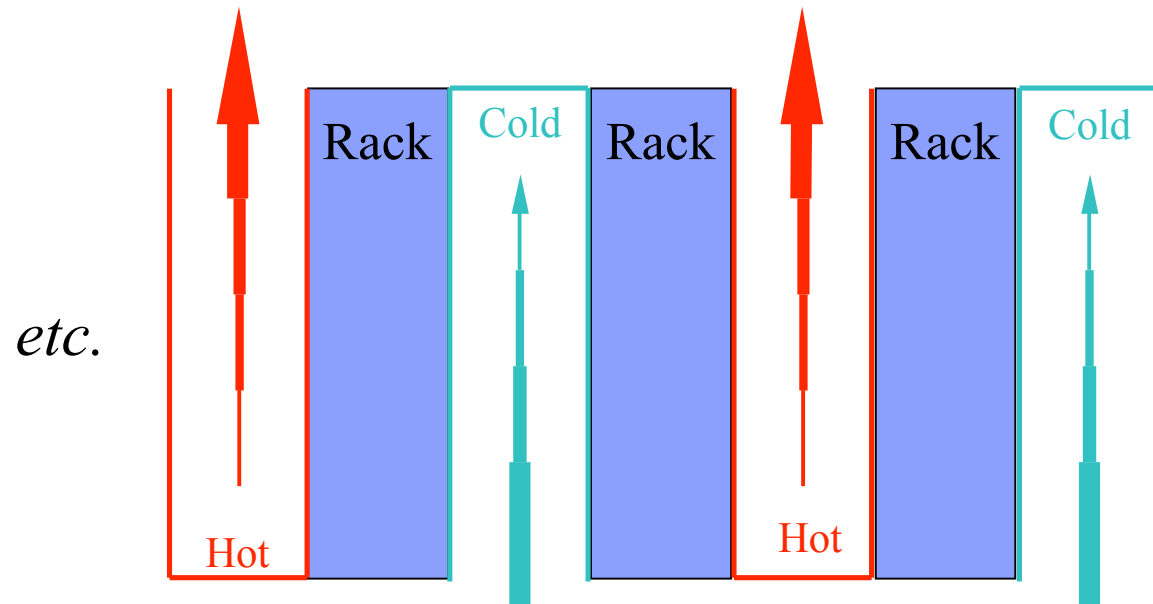
## First BG/P Rack



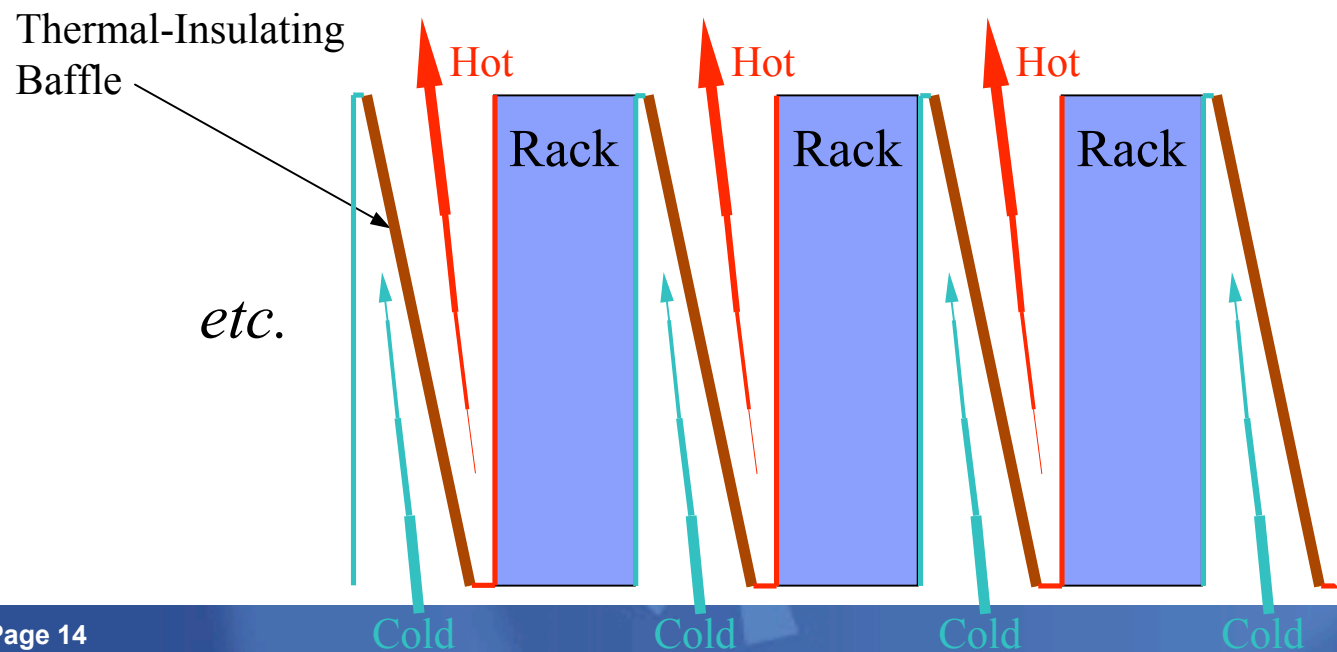


## First 8 racks of BG/P: Covers removed





**(a) Prior Art:  
Segregated,  
Non-Tapered  
Plenums**  
(Plenum Width Same  
Regardless of Flow Rate)



**(b) Invention:  
Integrated,  
Tapered  
Plenums**  
(Plenum Width  
Larger where Flow  
Rate is Greater)

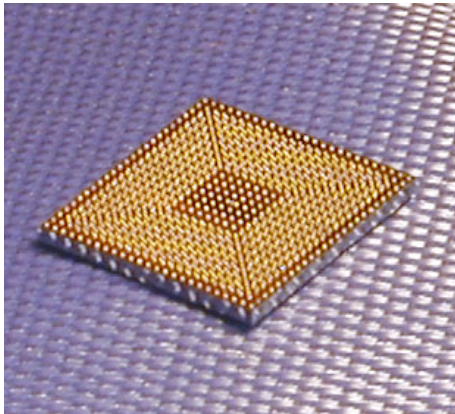
Shawn Hall 4-3-02  
02-04-03 Angled Plenums



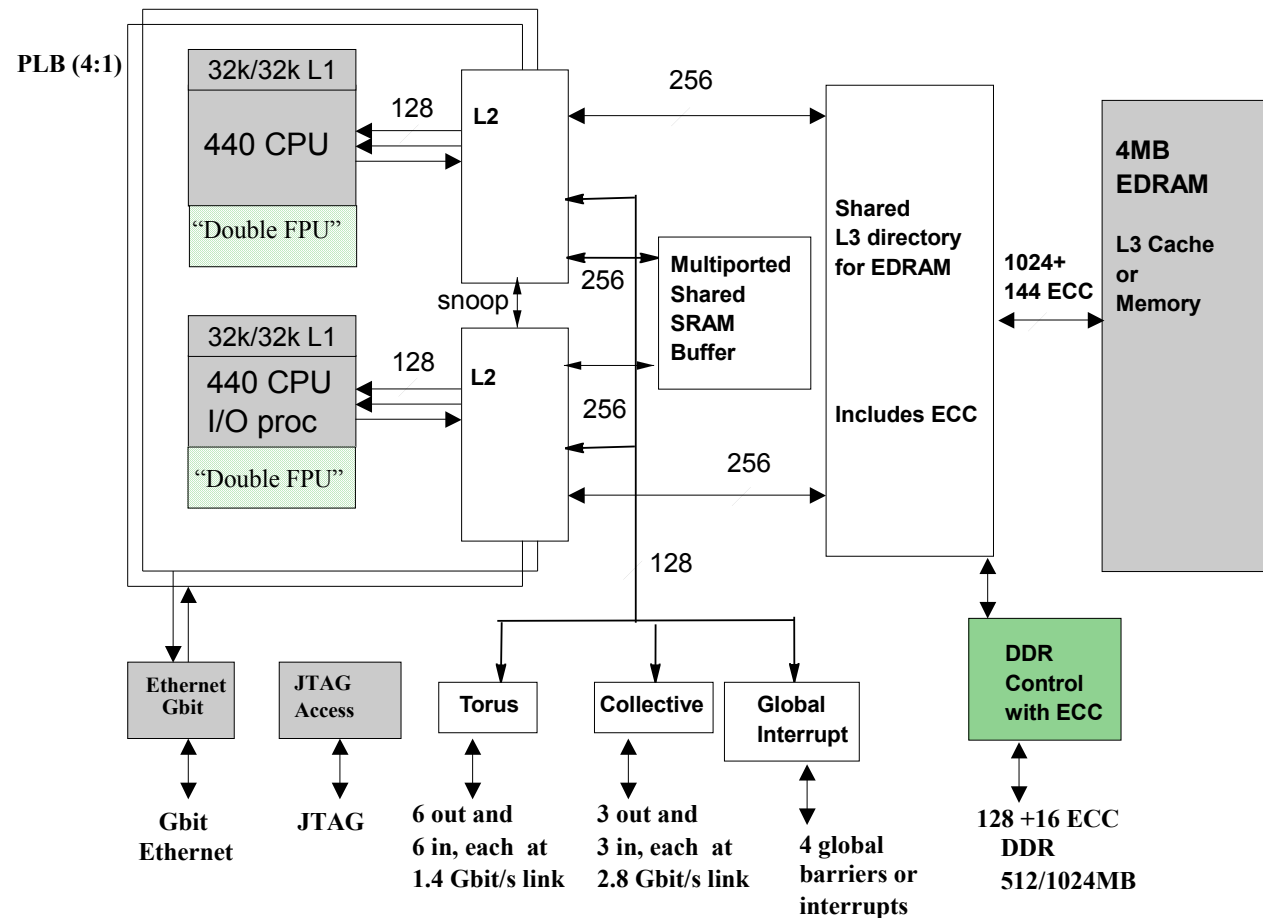
# IBM Blue Gene/P



# Blue Gene/L ASIC

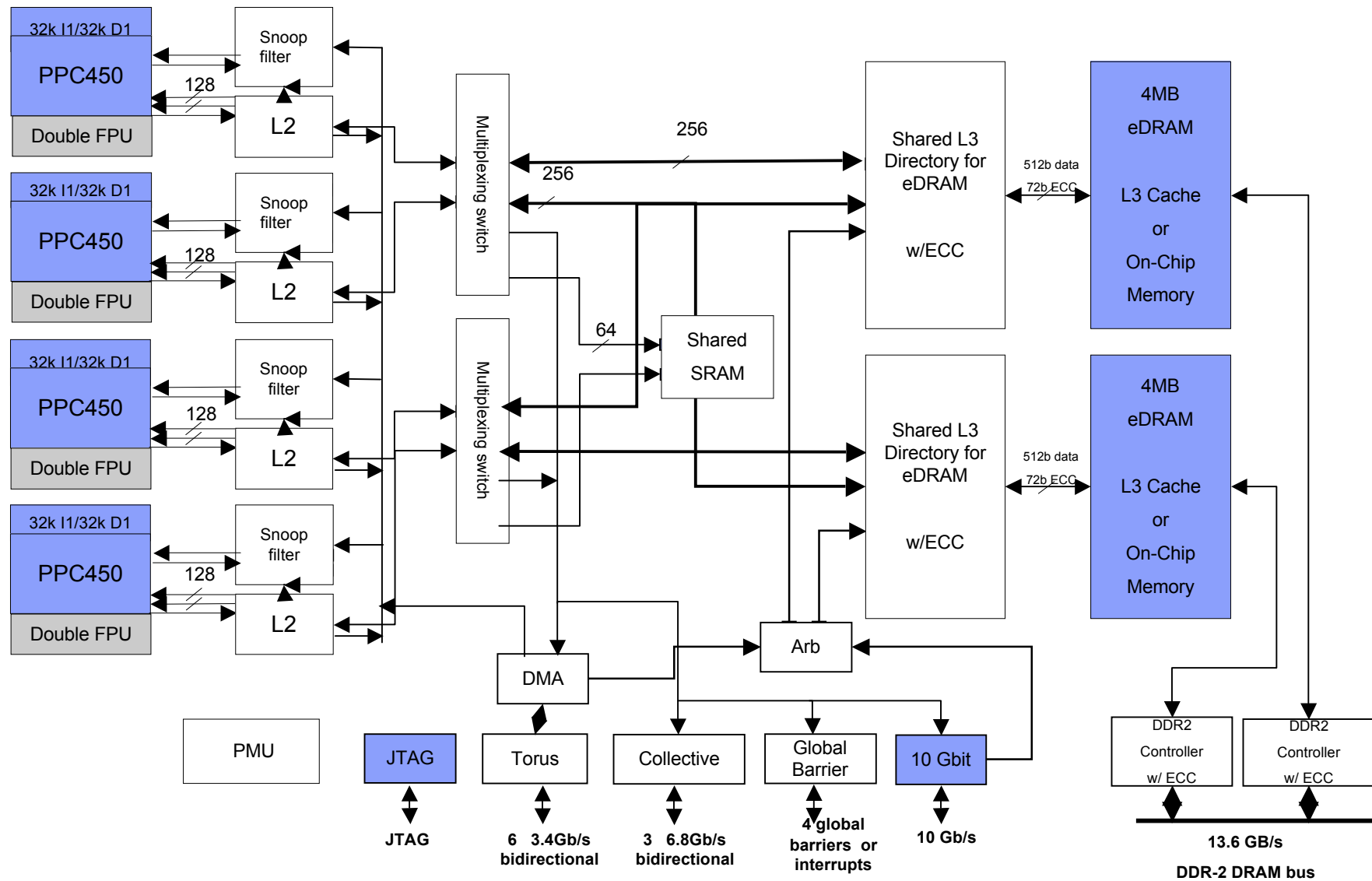


- IBM CU-11, 0.13  $\mu\text{m}$
- 11 x 11 mm die size
- 25 x 32 mm CBGA
- 474 pins, 328 signal
- 1.5/2.5 Volt

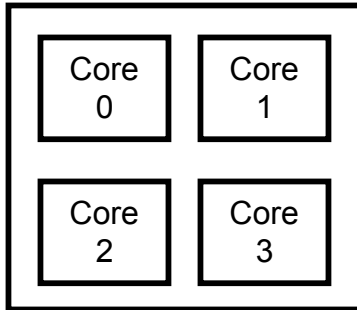




# Blue Gene/P ASIC

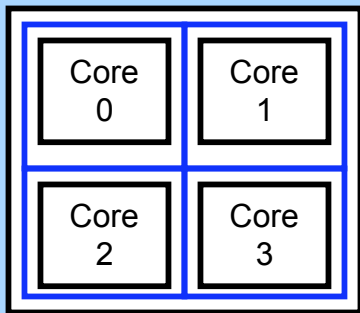


# Execution Modes in BG/P

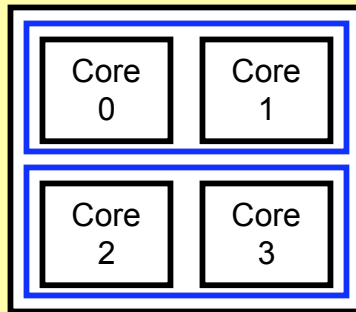


Hardware Elements Black  
Software Abstractions Blue

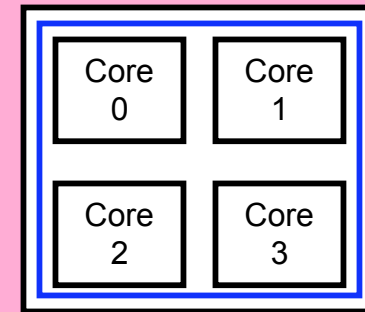
Quad Mode (VNM)  
4 Processes  
1 Thread/Process



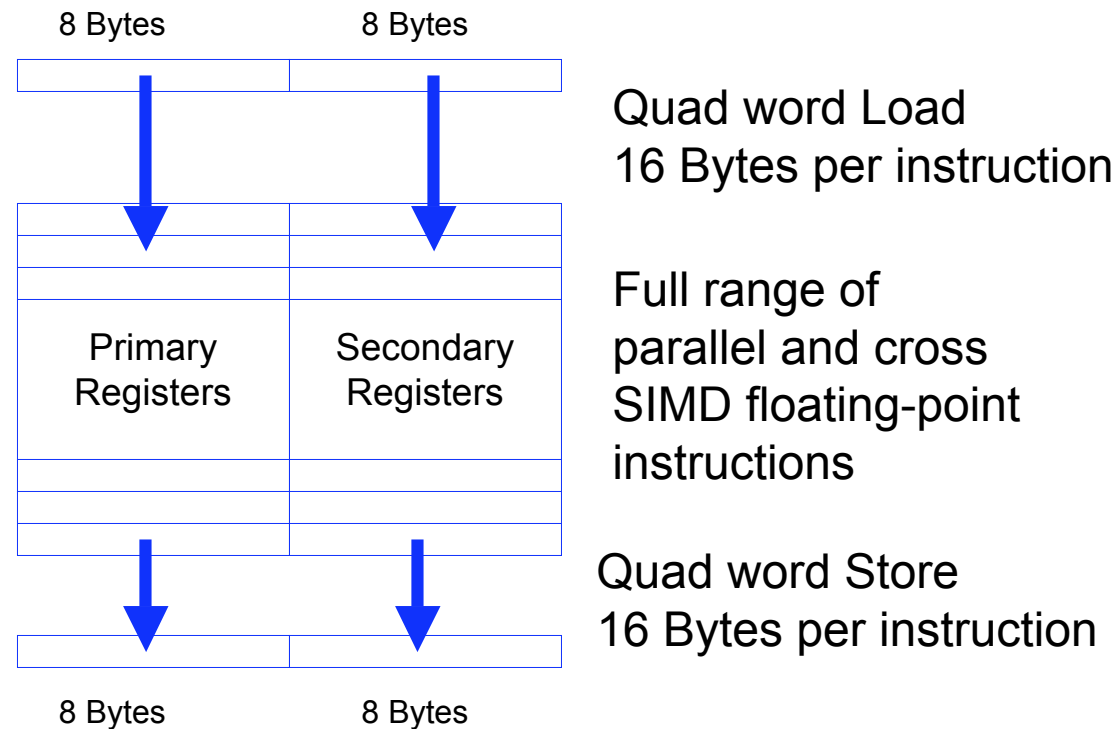
Dual Mode  
2 Processes  
1-2 Threads/Process



SMP Mode  
1 Process  
1-4 Threads/Process

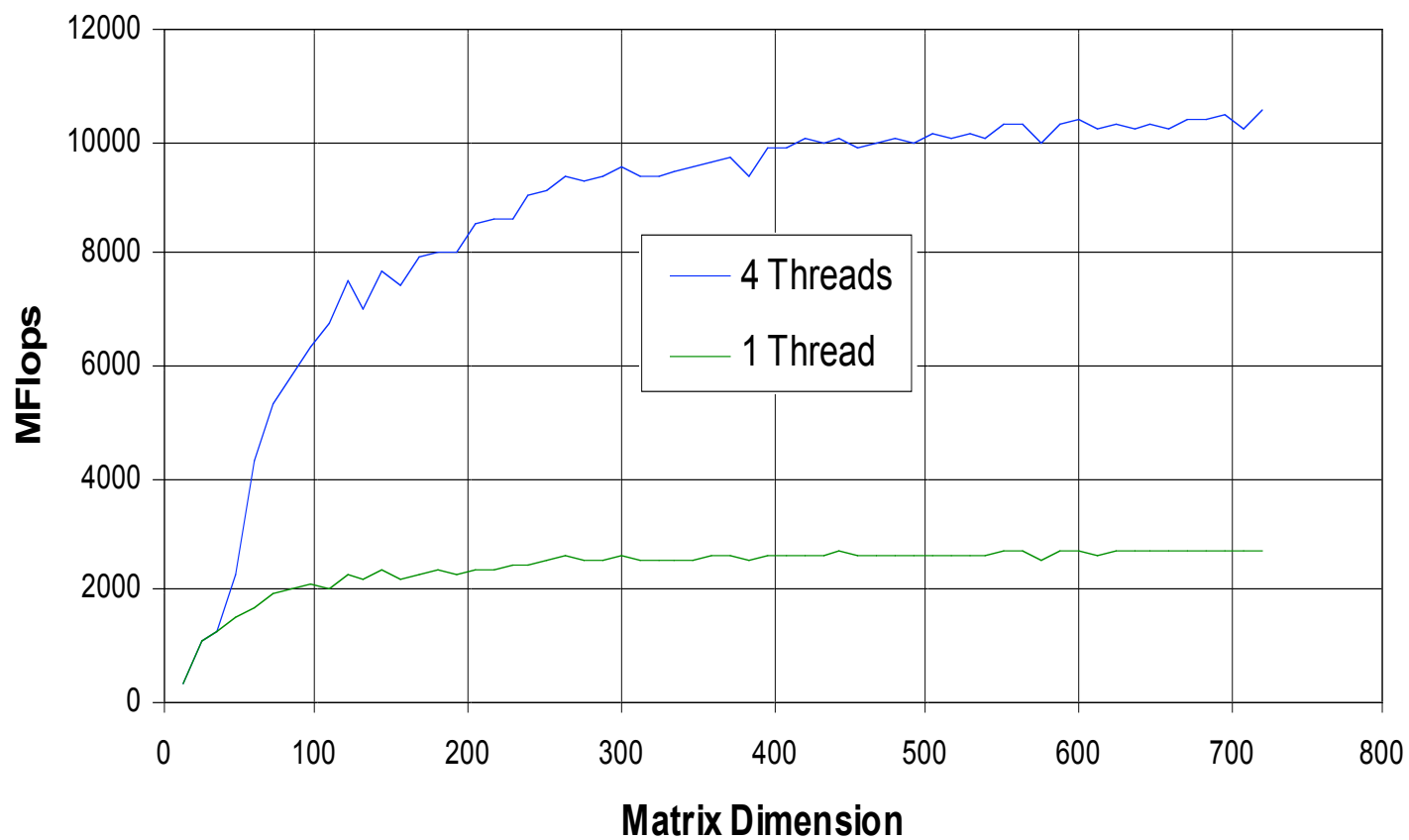


## Double Hummer Floating Point Unit

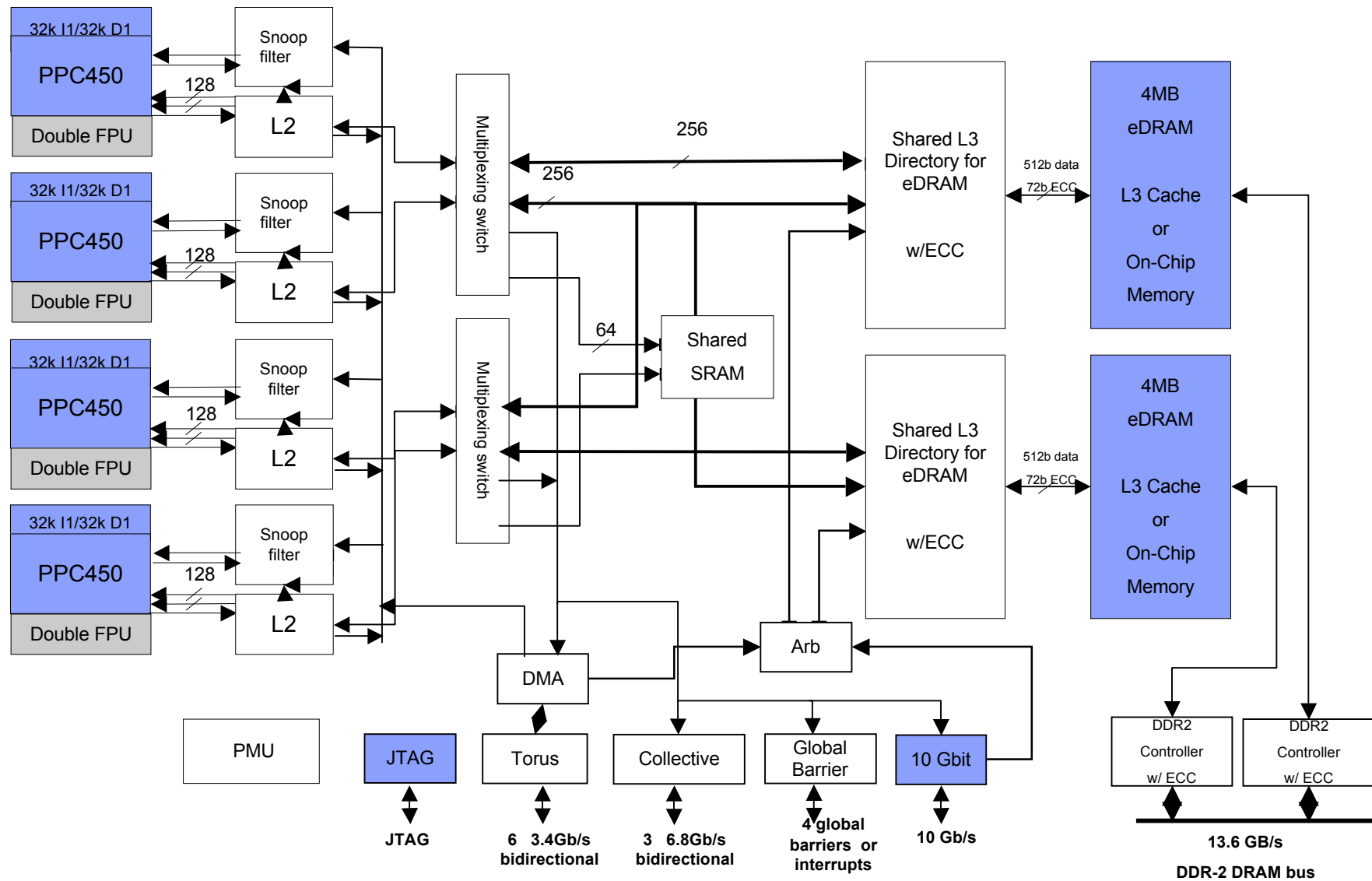


Quad word load/store operations  
require data aligned on 16-Byte  
boundaries.

## BG/P DGEMM (ESSL) One Node



# Blue Gene/P ASIC



# L1 Cache

## — Architecture

- 32KB Instruction-cache, 32KB Data-cache per core
- 32Byte lines, 64 way-set-associative, 16 sets
- Round-robin replacement
- Write-through mode
- No write allocation

## — Performance

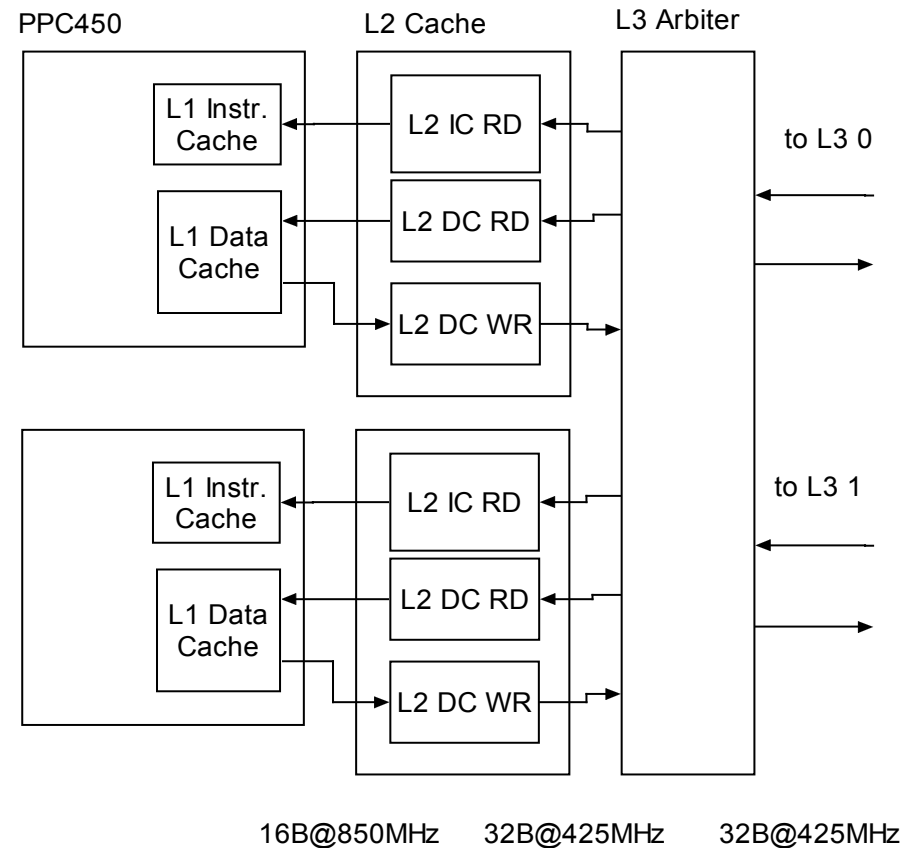
- L1 load hit => 8Bytes/cycle, 4 cycle latency (floating point)
- L1 load miss, L2 hit => 4.6Bytes/cycle, 12 cycle latency
- Store: Write through, limited by external logic to about one request every 2.9 cycles (about 5.5Bytes/cycle peak)

# L2 Cache

Three independent ports:  
Instruction read  
Data Read = prefetch  
Data write

## L3 Arbiter

Switch function  
1 read and 1 write  
switched to each L3  
every 425MHz cycle





# L2 Data Cache Prefetch Unit

Prefetch engine

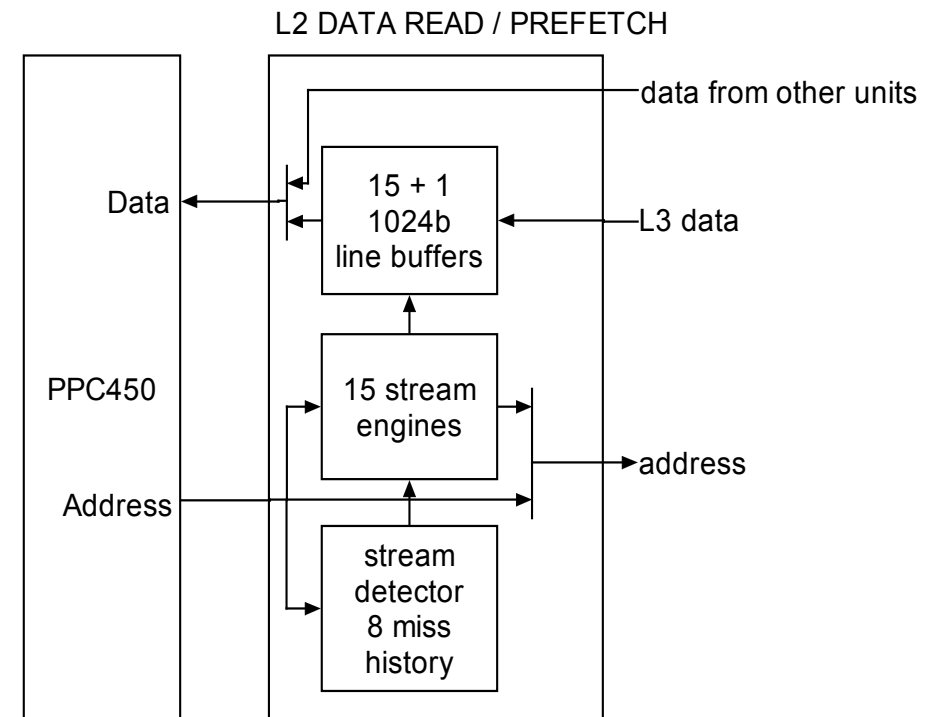
128Byte line prefetch

15 entry fully associative prefetch buffer

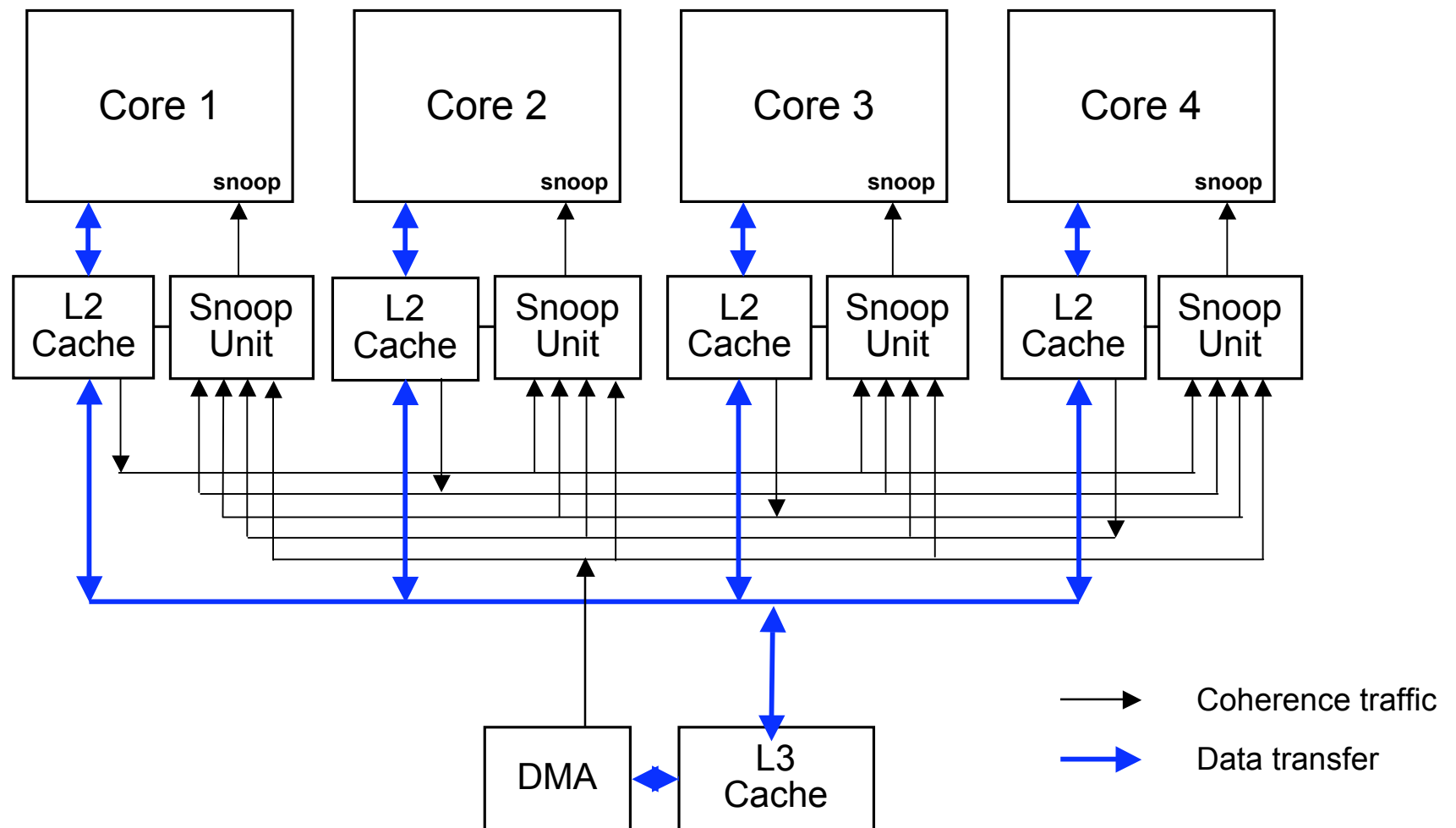
1 or 2 line deep prefetching

4.6Bytes/cycle, 12 cycle latency

Supports ~ 7 streams



## Snoop filters in the system

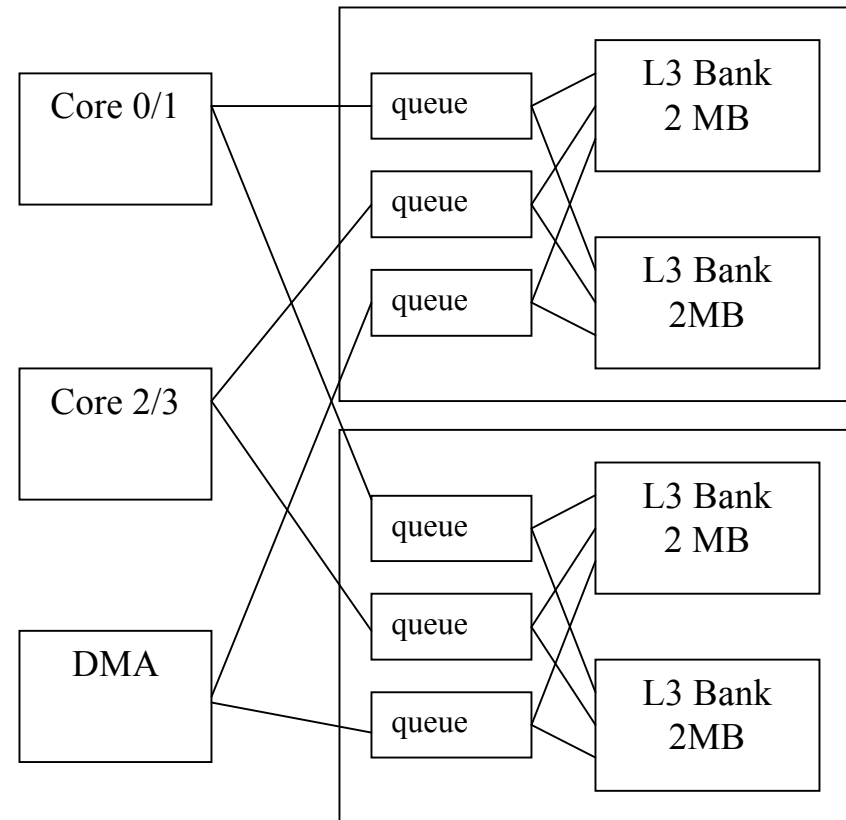


## L3 Cache

4 x 2 MB embedded DRAM  
banks per node (8MB total),  
each containing:

L3 directory

15 entry 128B-wide write  
combining buffer



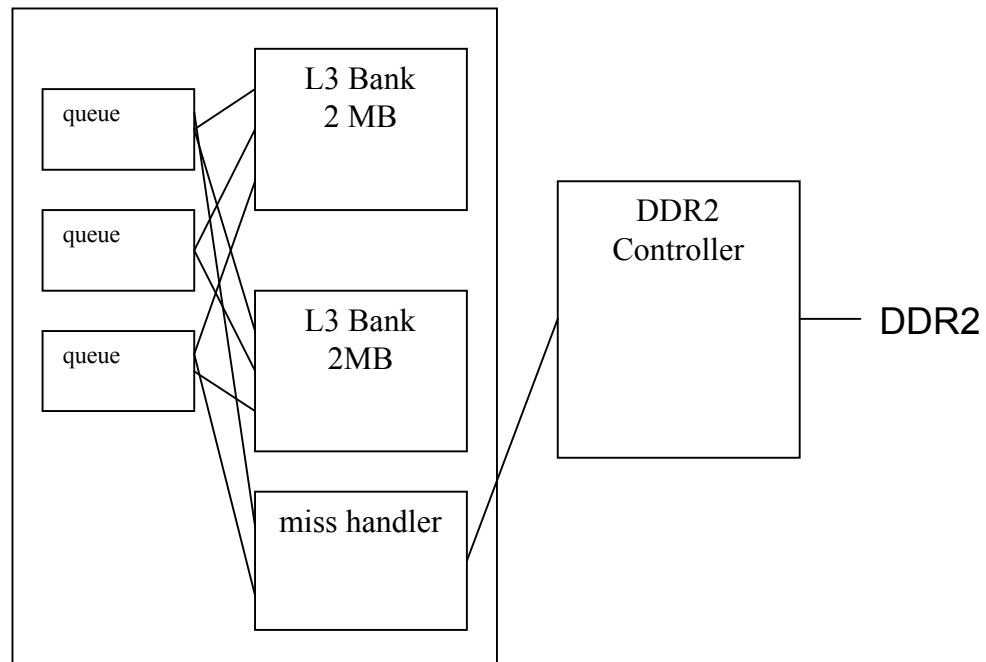
# Memory Controllers on Chip

20 8b-wide DDR2  
modules per controller

Controllers are on the  
chip

4 module-internal banks

Command reordering  
based on bank  
availability



# Memory System Bottlenecks

## L2 – L3 switch

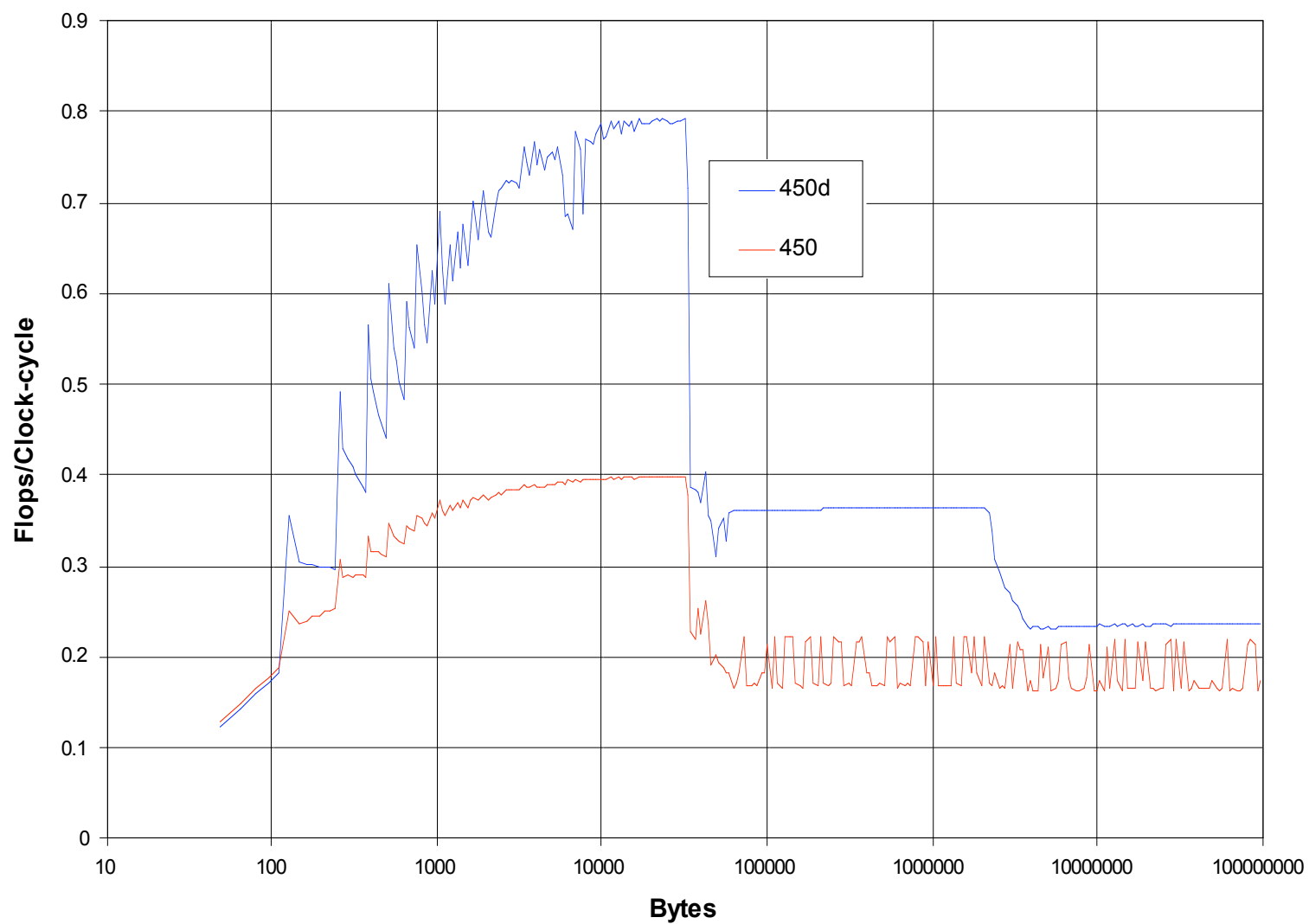
- Not a full core to L3 bank crossbar

- Request rate and bandwidth are limited if two cores of one dual processor group access the same L3 cache bank

## Banking for DDR2

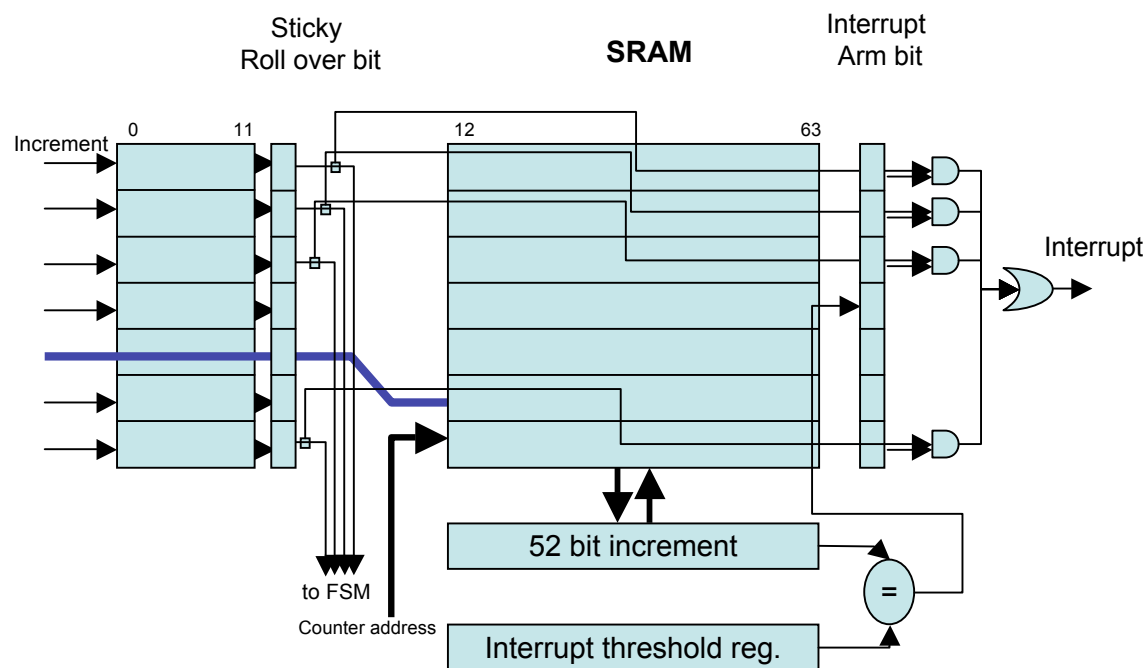
- 4 banks on 512Mb DDR modules

- Peak bandwidth only achievable if accessing 3 other banks before accessing the same bank again

**BGP Daxpy Performance**

# Performance Monitor Architecture

- Novel hybrid counter architecture
  - High density and low power using SRAM design
- 256 counters with 64bits resolution
  - Fast interrupt trigger with configurable threshold
  - Performance analysis is key to achieving full system potential





# Performance Monitor Features

- Counters for core events

loads, stores, floating-point operations (flops)

- Counters for the memory subsystem

cache misses, DDR traffic, prefetch info, etc.

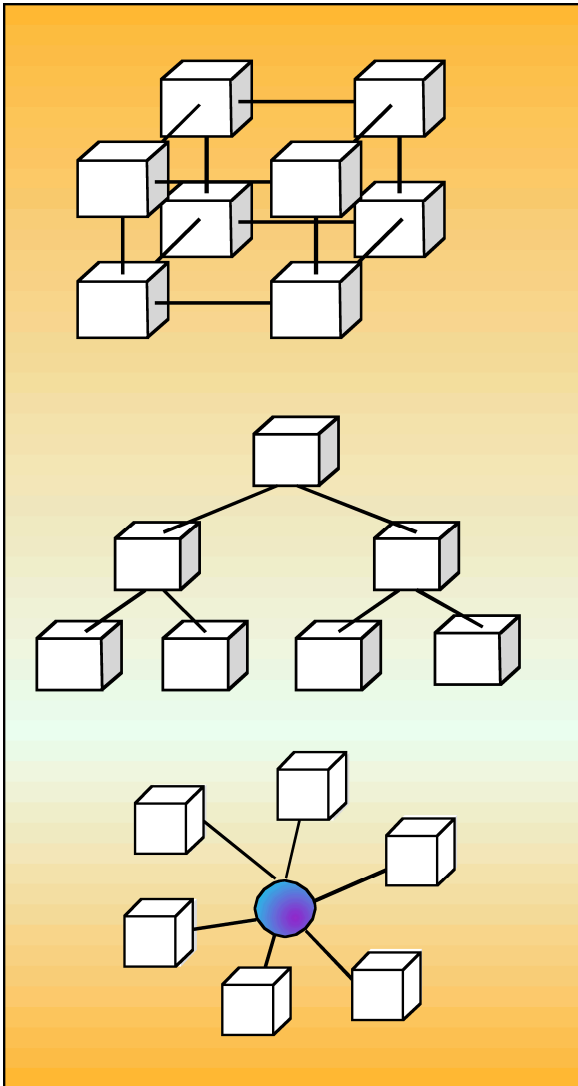
- Counters for the network interfaces

torus traffic, collective network, DMA, ...

- Counts are tied to hardware elements

counts are for cores or nodes, not processes or threads

# Blue Gene/P Interconnection Networks



## 3 Dimensional Torus

- Interconnects all compute nodes
- Virtual cut-through hardware routing
- 3.4 Gb/s on all 12 node links (5.1 GB/s per node)
- 0.5  $\mu$ s latency between nearest neighbors, 5  $\mu$ s to the farthest
- MPI: 3  $\mu$ s latency for one hop, 10  $\mu$ s to the farthest
- Communications backbone for point-to-point

## Collective Network

- One-to-all broadcast functionality
- Reduction operations for integers and doubles
- 6.8 Gb/s of bandwidth per link per direction
- Latency of one way tree traversal 1.3  $\mu$ s, MPI 5  $\mu$ s
- Interconnects all compute nodes and I/O nodes

## Low Latency Global Barrier and Interrupt

- Latency of one way to reach 72K nodes 0.65  $\mu$ s, MPI 1.6  $\mu$ s

# Blue Gene/P Torus Network

Logic Unchanged from BG/L, *except*

## Bandwidth

BG/L: clocked at  $\frac{1}{4}$  processor rate      1Byte per 4 cycles

BG/P: clocked at  $\frac{1}{2}$  processor rate      1Byte per 2 cycles

With frequency bump from 700 MHz to 850 MHz

BG/P Links are 2.4x faster than BG/L

**425** MB/s vs **175** MB/s

Same Network Bandwidth per Flops as BG/L

Minor changes in error reporting

Primary interface is via DMA, rather than cores

Run application in DMA mode, or core mode (not mixed)

Software product stack uses DMA mode

# DMA Overview

Rich function DMA offloads data movement to/from torus

## Message Types

Direct Put:	copy from source address to destination address
Remote Get:	remote copy from target node address to source
Memory Fifo:	place packets in Fifo (in memory) on destination

Arbitrary offsets

Can do memcpy within a node

## Basic Constructs:

- Injection and Reception Fifos

- Injection and Reception “Counters”

- Organized into 4 “Groups” (could have one group/core)

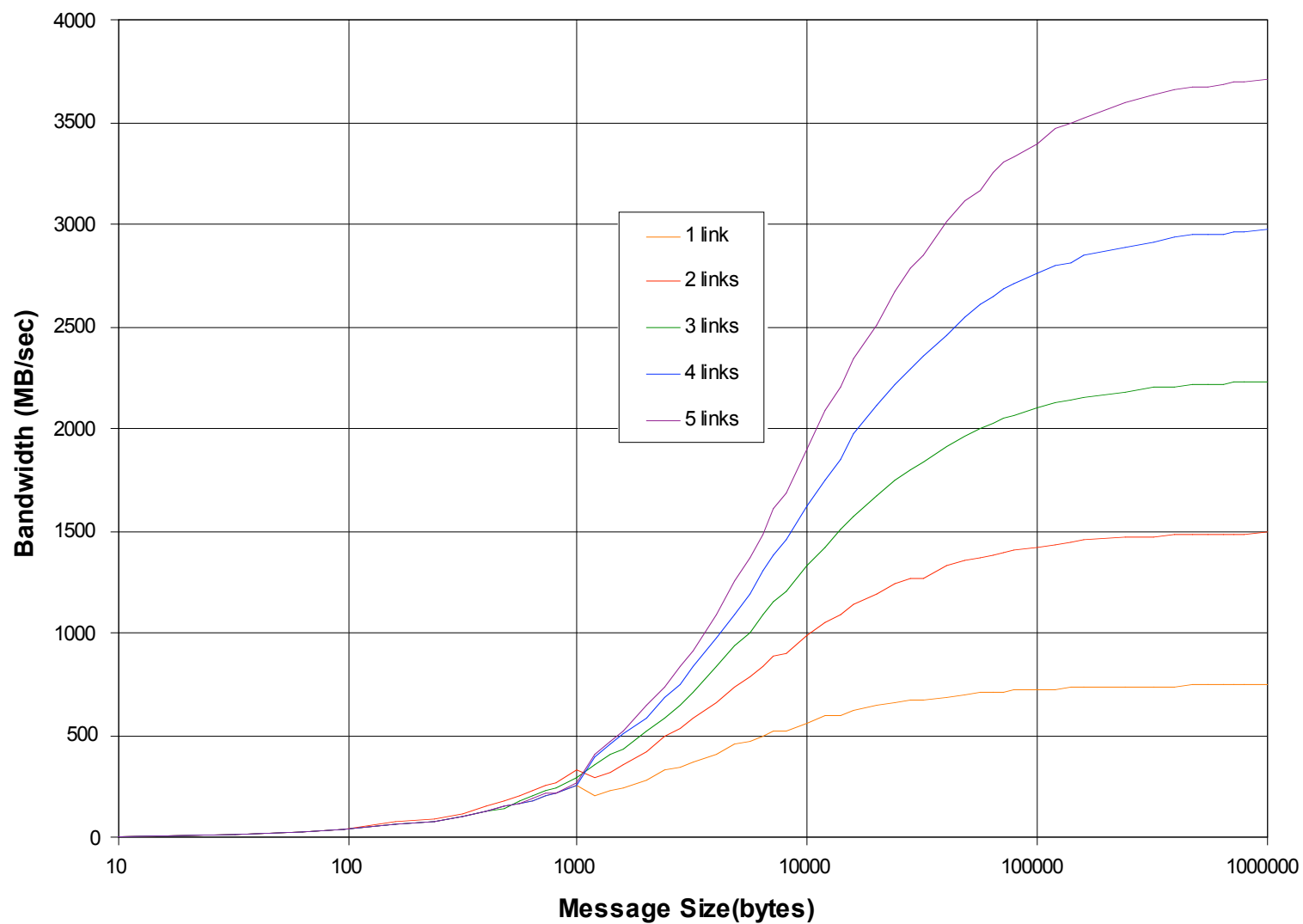
- Message Descriptors (32 B)

  - Includes destination, message type and length, payload pointer

DMA uses physical addresses

- Efficient SPI programming interface for Virtual to Real translation

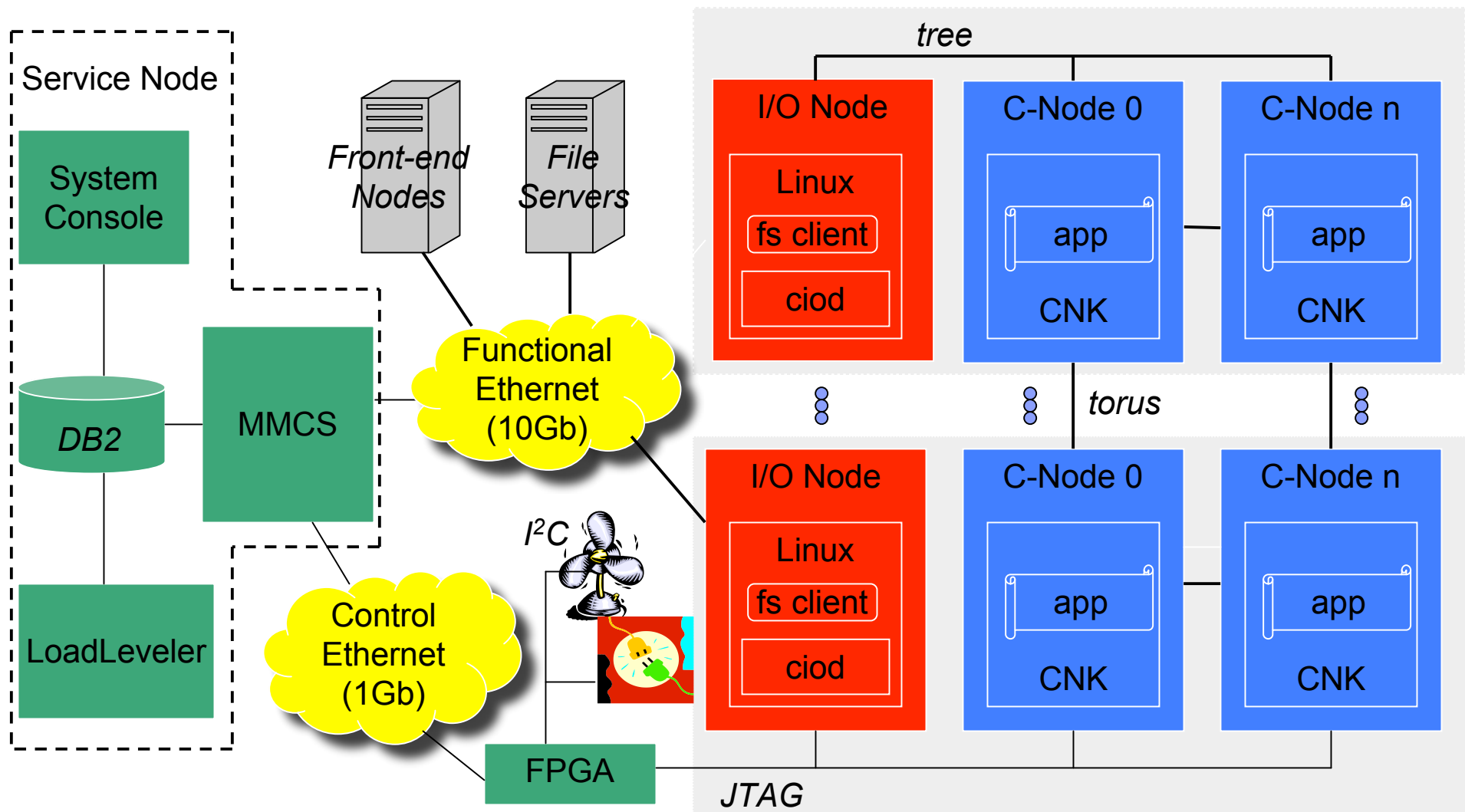
### BGP Exchange Bandwidth



## Blue Gene/P key architectural improvements over BG/L

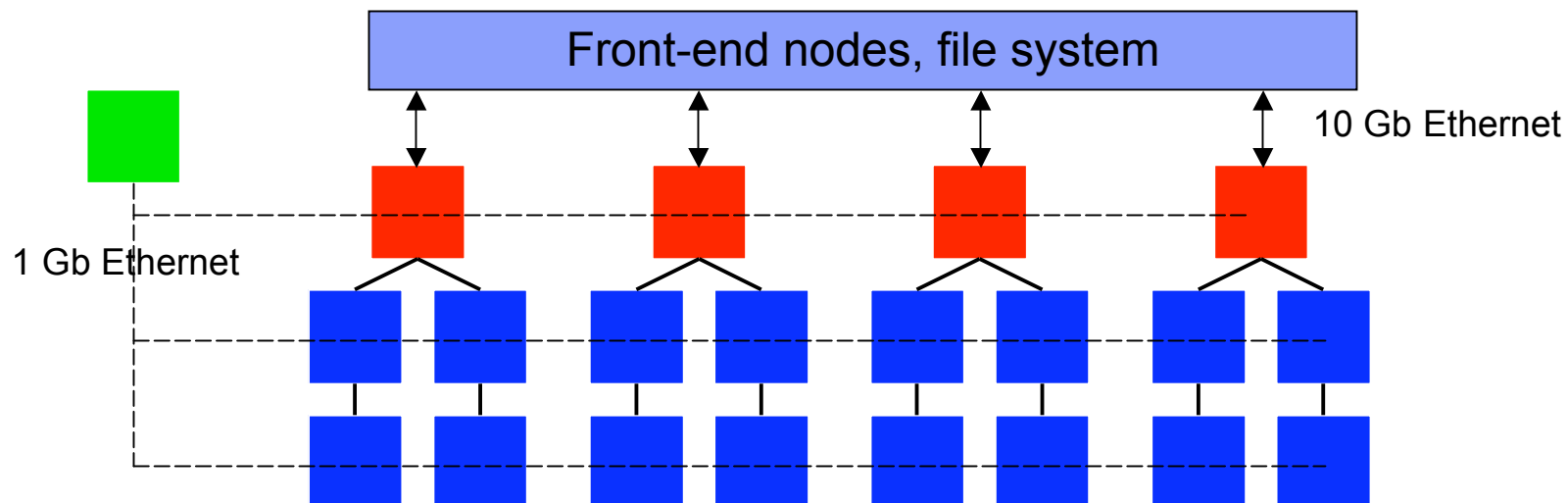
Property		Blue Gene/L	Blue Gene/P
Node Properties	Processor cores/chip	two 440 PowerPC	four 450 PowerPC
	Processor Frequency	0.7GHz	0.85GHz
	Coherency	Software managed	SMP with snoop filtering
	L1 Cache (private)	32KB I + 32KB D/proc.	32KB I + 32KB D/proc.
	L2 Cache (private)	15 line buffers	15 line buffers
	L3 Cache size (shared)	4MB	8MB
	Main Store	512 MB and 1 GB DDR	2 GB DDR2
	Main Store Bandwidth	5.6 GB/s (16B wide)	13.6 GB/s (2*16B wide)
	Peak Performance	5.6 GFLOPs/node	13.6 GFLOPs/node
Torus Network	Aggregate Bandwidth	6*2*175 MB/s=2.1 GB/s	6*2*425 MB/s= 5.1 GB/s
	Hardware Latency (Nearest Neighbor)	<1μs	<1μs
Collective Network	Aggregate Bandwidth	3*2*350 MB/s=2.1 GB/s	3*2*0.85 GB/s=5.1 GB/s
Performance Monitors	Counters	48	256
	Counter resolution (bits)	32b	64b
GFLOPS/Watt		0.22	0.33

# Blue Gene/P System Architecture



# Blue Gene Software Hierarchical Organization

- **Compute nodes** dedicated to running user application, and almost nothing else - simple compute node kernel (CNK)
- **I/O nodes** run Linux and provide a more complete range of OS services – files, sockets, process launch, signaling, debugging, and termination
- **Service node** performs system management services (e.g., partitioning, heart beating, monitoring errors) - transparent to application software



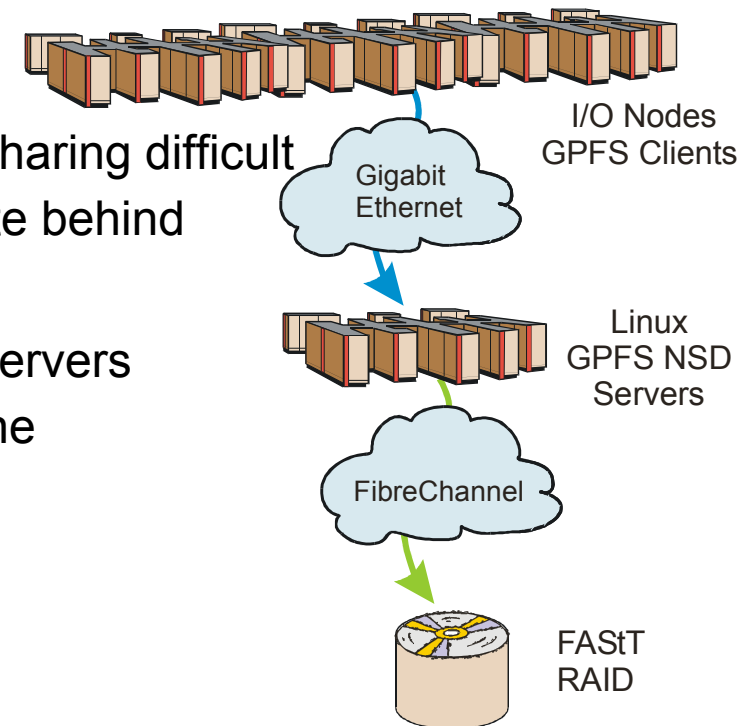


# Programming models and development environment

- Familiar methods
  - SPMD model - Fortran, C, C++ with MPI (MPI1 + subset of MPI2)
    - Full language support with IBM XL and GNU compilers
    - Automatic SIMD FPU exploitation (limited)
  - Linux development environment
    - User interacts with system through front-end nodes running Linux – compilation, job submission, debugging
    - Compute Node Kernel provides look and feel of a Linux environment
      - POSIX routines (with some restrictions: no fork() or system())
      - BG/P adds pthread support, additional socket support
    - Tools – support for debuggers, MPI tracer, profiler, hardware performance monitors, visualizer (HPC Toolkit), PAPI
- Restrictions (which lead to significant benefits)
  - Space sharing - one parallel job per partition of machine, one thread per core in each compute node
  - Virtual memory is constrained to physical memory size

# General Parallel File System (GPFS) for Blue Gene

- Blue Gene can generate enormous I/O demand (disk limited)
  - BG/P IO-rich has 64 10Gb/rack – 80GB/sec
- Serving this kind of demand requires a parallel file system
- NFS for file I/O
  - Limited scalability
  - NFS has no cache consistency, making write sharing difficult
  - Poor performance, not enough read ahead/write behind
- GPFS runs on Blue Gene
  - GPFS clients in Blue Gene call external NSD servers
  - Brings traditional benefits of GPFS to Blue Gene
    - I/O parallelism
    - Cache consistent shared access
    - Aggressive read-ahead, write-behind



# BG/P Single Node Software Enhancements

- Compute Node Kernel
  - SMP support – multithreading
  - Enhancements to support broader class of applications on-demand linking, I/O services
  - Support for enhanced virtual node mode with sharing of read-only data
- I/O node kernel
  - Full SMP support
  - Drive 10 Gb ethernet
- Compilers
  - OpenMP support to version 2.5 specification
- Math libraries
  - ESSL, MASS, MASSV – complete functionality

# BG/P Multi-node Software Enhancements

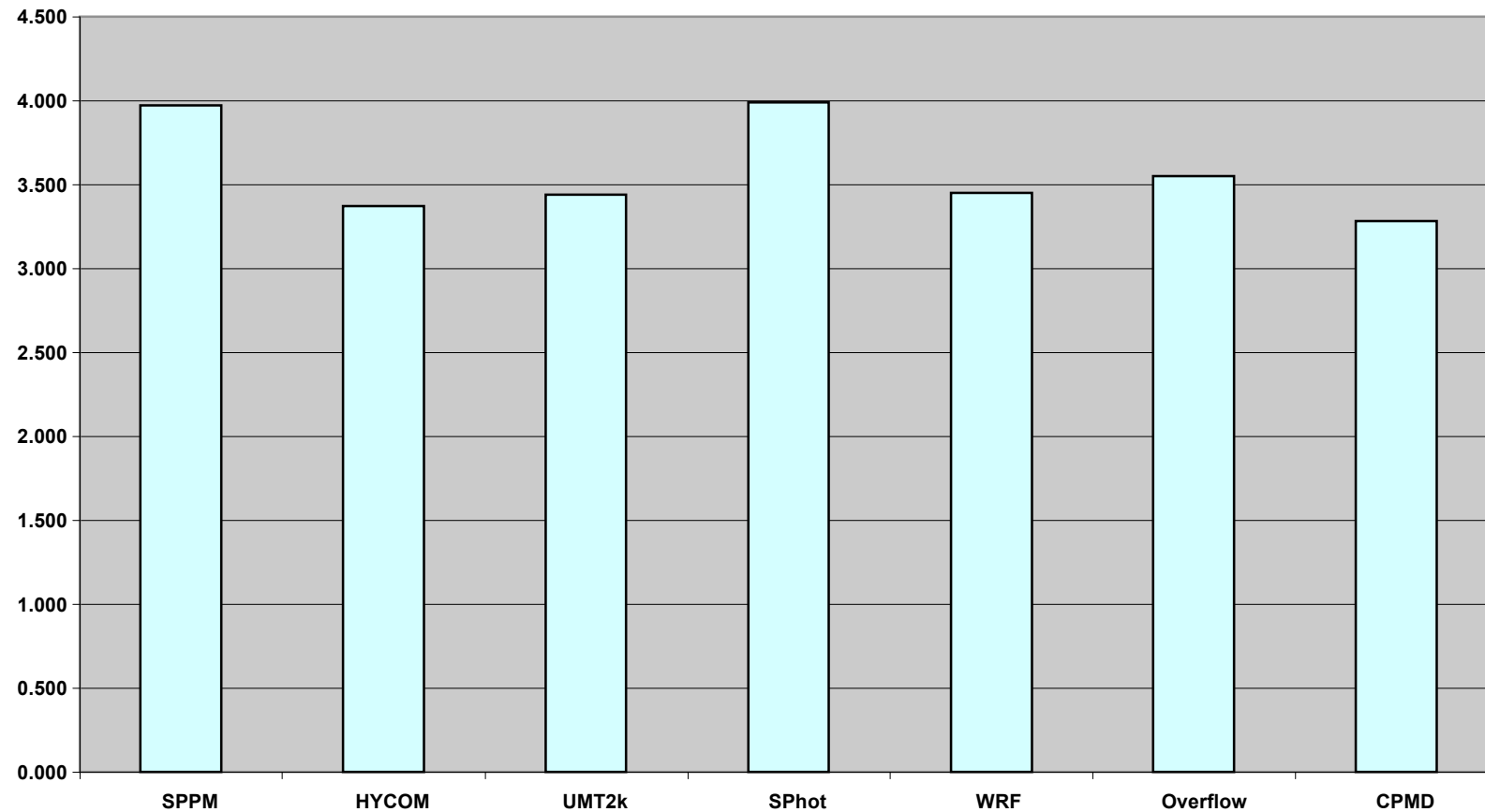
- MPI
  - Full MPI2 support, except for dynamic process management
    - MPI-IO optimizations, one-sided communication
  - Exploit DMA for the torus network
  - Exploit multiple processors for single MPI call
  - Higher levels of scalability
    - Adaptive protocol selection
    - Adaptive connection and buffer management
- Control system
  - Enhanced ease of management
  - Higher levels of scalability
  - Proactive handling of failures
- Job scheduler/launch
  - Integration in a grid environment – making Blue Gene available as a managed web service

# BG/P Linpack Performance

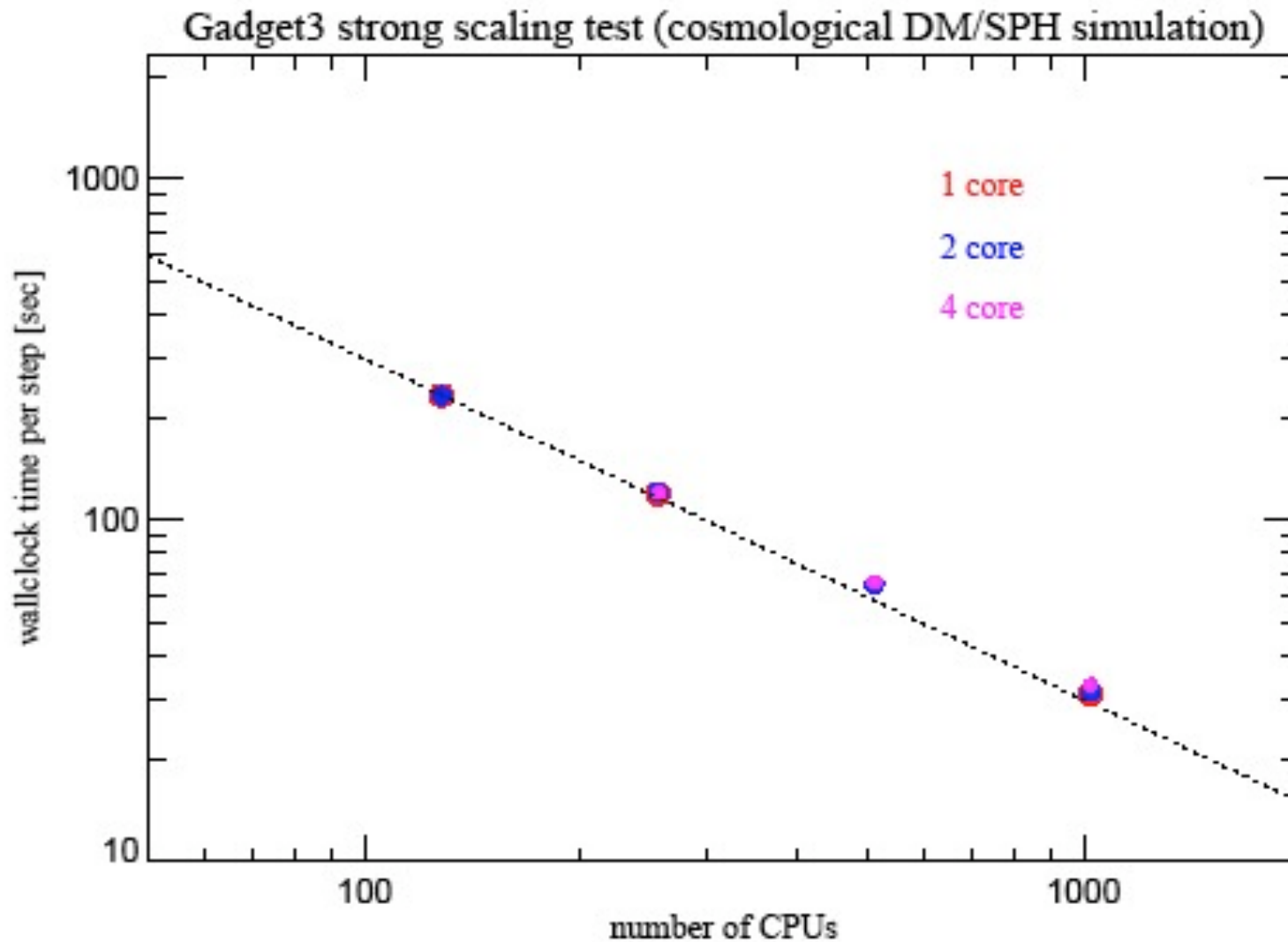
<b>System Size</b>	<b>Nodes</b>	<b>Processors</b>	<b>Performance (TF)</b>	<b>Fraction of Peak</b>
<b>1 Rack</b>	<b>1024</b>	<b>4096</b>	<b>11.10</b>	<b>80%</b>
<b>2 Racks</b>	<b>2048</b>	<b>8192</b>	<b>21.91</b>	<b>79%</b>
<b>4 Racks</b>	<b>4096</b>	<b>16384</b>	<b>43.96</b>	<b>79%</b>
<b>8 Racks</b>	<b>8192</b>	<b>32768</b>	<b>85.98</b>	<b>77%</b>
<b>16 Racks</b>	<b>16384</b>	<b>65536</b>	<b>(Top 500) 167.20</b> <b>(ANL Dec 13)</b> <b>171.80</b>	<b>75%</b> <b>77%</b>

# OpenMP Scaling

Misc codes: OMP Speedup on 4 threads wrt a single thread



# Strong Scaling of Applications on BG/P



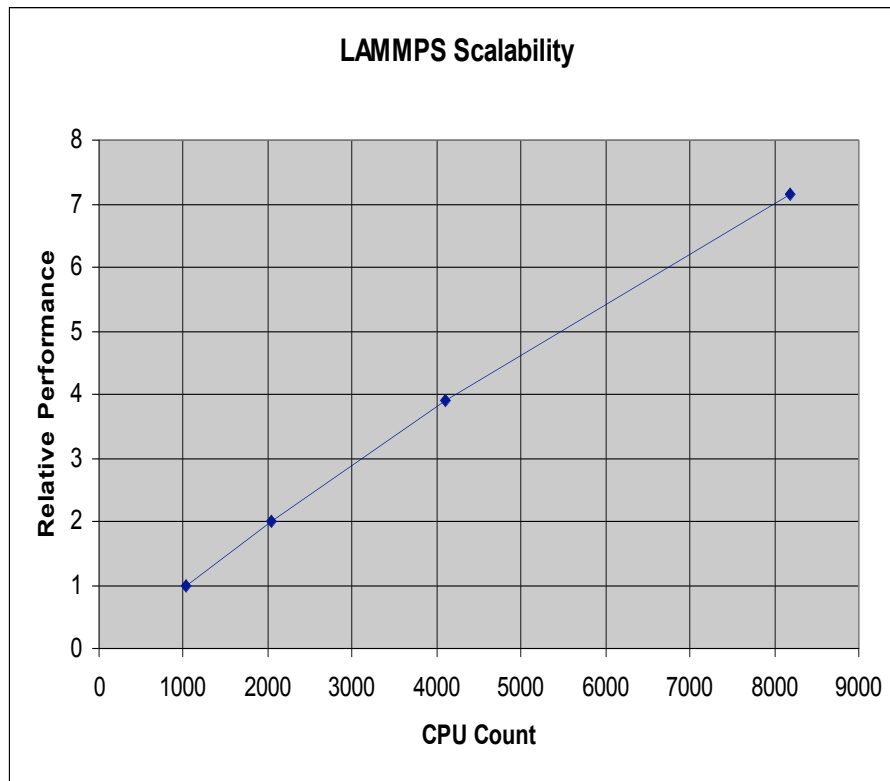
Courtesy  
of  
Volker  
Springel

Max  
Planck  
Institute  
for  
Astro-  
physics  
Garching

# Strong Scaling of Applications on BG/P

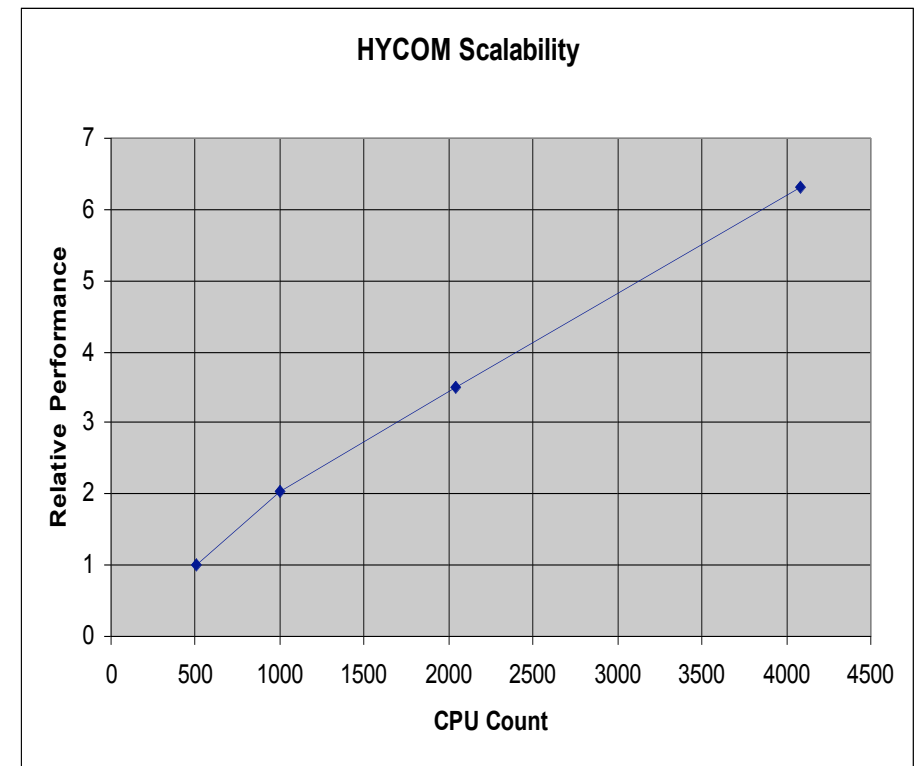
## ■ LAMMPS

- Molecular Dynamics
- Strong Scaling Problem (fixed problem size)



## ■ HYCOM

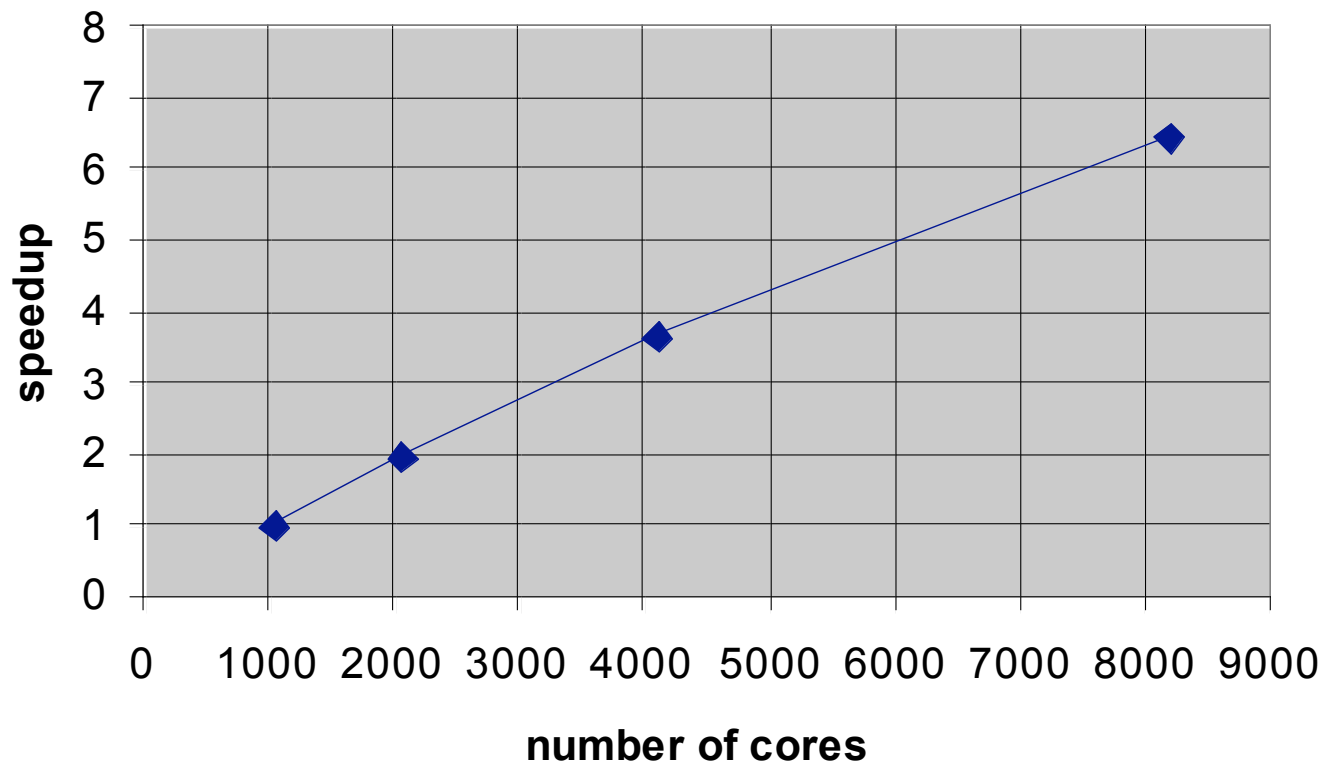
- Ocean Model
- Strong Scaling Problem (fixed problem size)





# GENE Scaling on BlueGene/P

strong scaling, problem size 0.5 TB  
using 4 cores per node

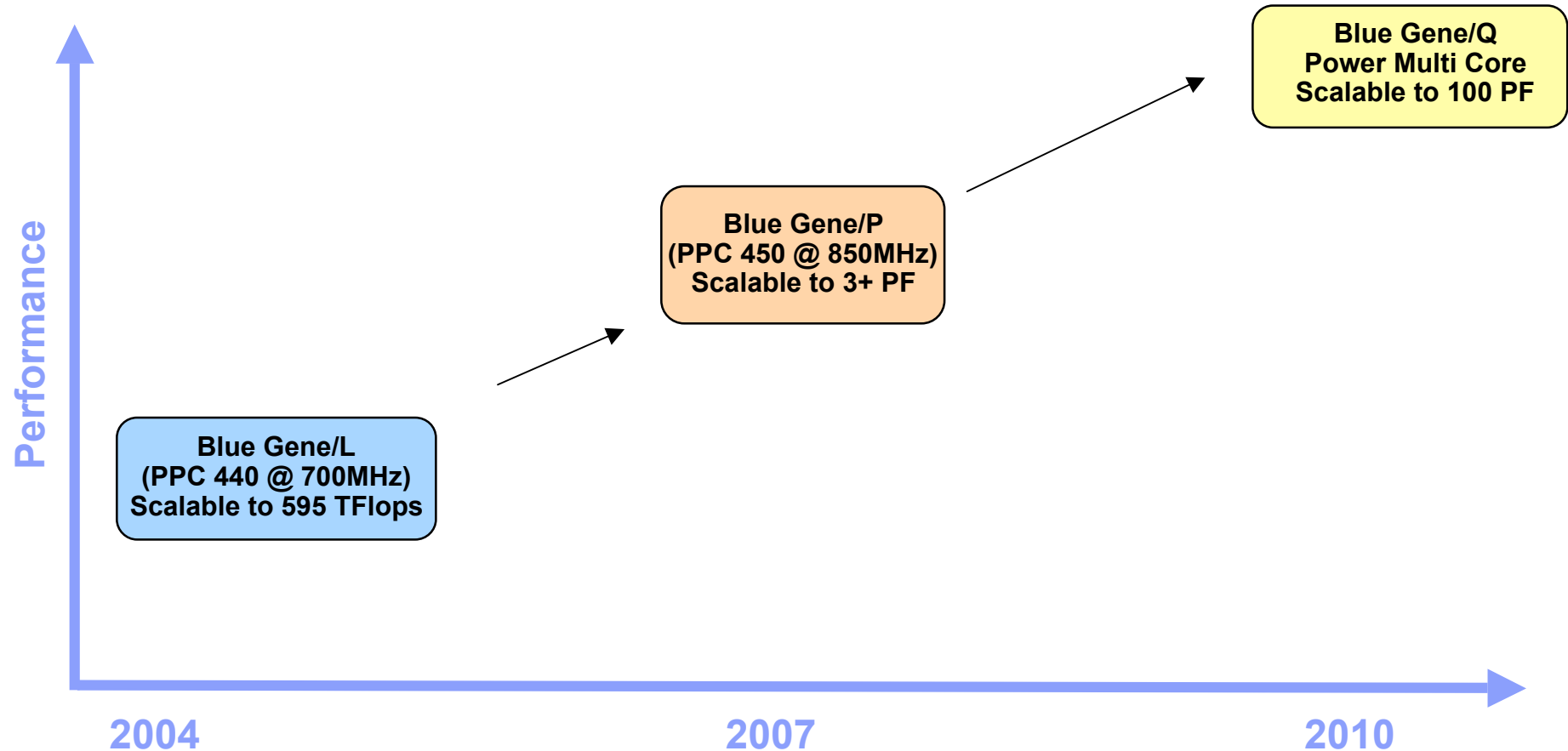


Gyrokinetic  
Electromagnetic  
Numerical  
Experiment  
Max Planck

# Summary

- Blue Gene/P: Facilitating Extreme Scalability
  - Ultrascale capability computing
  - Provides customer with enough computing resources to help solve grand challenge problems
  - Provide competitive advantages for customers' applications
  - Energy conscious solution supporting green initiatives
  - Familiar open/standards operating environment
  - Simple porting of parallel codes
- Key Solution Highlights
  - Leadership performance, space saving design, low power requirements, high reliability, and easy manageability

# Blue Gene Technology Roadmap



Note: All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.